CIVIC DIGITAL FELLOWSHIP

Optimizing the Commodity Flow Survey (CFS) with Machine Learning

Alex Gao

Supervised by Christian Moscardi

Economic Reimbursable Surveys Division

Problem: non-shipping establishments use up CFS resources

"We don't ship." Are they reporting accurately?

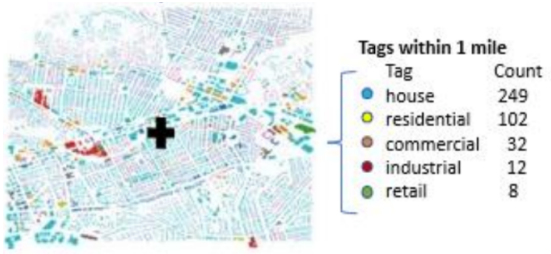
- Manually verify whether an establishment ships
- Inspect satellite imagery, data about the address, etc
- Contact the establishment and ask for clarification
- Send them the survey again next quarter

Approx. $450,000 spent each CFS on establishments that should be out-of-scope of the survey in the first place.

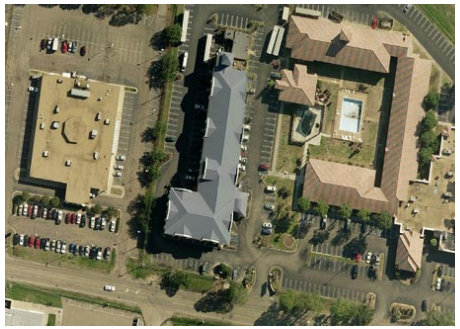
Solution: use machine learning to preemptively detect non-shipping locations

Existing tools (built in 2019) to collect metadata from an address

Openstreetmaps data pipeline

(lat, lon) → .osm files →  → 

Satellite image retrieval

(lat, lon) → Census geo division →  → 

Note: all data here is dummy data

Goals:

1. Combine and expand existing datasets
2. Improve model predictions
3. Filter out non-shipping locations, saving time and money

Challenges and new problems

Fixing old pipelines

- Undetected bugs that led to the selection and calculation of incorrect data
- Slow and prohibitive to scale and experiment with
- **Solution: better testing and debugging, and parallelize**

Limited dataset

- Only ~25,000 entries to train on from the CFS data
- The model needs more non-shipping locations to train on
- **Solution: incorporate Business Register data**

Messy sources of truth

- Addresses labeled as both shipping and nonshipping, given different coordinates
- Conflicts between Business Register and CFS data
- **Solution: investigation and coordination with others**

Deliverables and future work

1

New and improved model

- **90% accuracy** on test data that was hidden from the model during training
- Improvement over past model with **78% accuracy**

2

Documented and reported data discrepancies to appropriate teams

- Created notebook demonstrating inaccurate geocoding
- Over **10%** of addresses in the Business Register had geocoding **off by >1 kilometer**

3

Extrapolated model to Business Register

- Re-geocoded as many locations as possible; tested model on that data
- This is a necessary step to inform future CFS sampling frames

Future: properly integrate expanded dataset into the model; apply similar methods to identify construction activity, retail activity, etc